STATISTICS IN MEDICINE
Statist. Med. (in press)
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2562

Comparison of validity of assessment methods using indices of adjusted agreement

Jun-mo Nam[*,†]

Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Department of Health & Human Services, Executive Plaza South, Room 8028, 6120 Executive Boulevard, MSC 7240, Rockville, MD 20892-7240, U.S.A.

SUMMARY

For comparing the validity of rating methods, the adjusted κ (S coefficient) and Yule's Y index are better than Cohen's κ which is affected by marginal probabilities. We consider a validity study in which a subject is assessed as exposed or not-exposed by two competing rating methods and the gold standard. We are interested in one of the methods, which is closer in agreement with the gold standard. We present statistical methods taking correlations into account for comparing the validity of the rating methods using S coefficient and Y index. We show how the S coefficient and Yule's Y index are related to sensitivity and specificity. In comparing the two rating methods, the preference is clear when the inference is the same for both S and Y. If the inference using S differs from that using Y, then it is not obvious how to decide a preference. This may occur when one rating method is better than the other in sensitivity but not in specificity. Numerical examples for comparing asbestos-exposure assessment methods are illustrated. Copyright © 2006 John Wiley & Sons, Ltd.

KEY WORDS: adjusted agreement; comparison of raters; sensitivity; specificity; Yule's index

1. INTRODUCTION

Cohen's κ [1] has been commonly applied in the evaluation of interrater agreement. Interval estimation of Cohen's κ agreement has been studied by various authors, e.g. Fleiss *et al.* [2], Fleiss and Cicchetti [3], and Hale and Fleiss [4]. Despite its popularity, there has been criticism of the κ statistic. A number of authors have pointed out paradoxes associated with the effect of marginal proportions, e.g. References [5, 6]. The κ is affected in complex ways by the bias between raters and by prevalence. For a resolution of these paradoxes, Byrt *et al.*

[*] Correspondence to: Jun-mo Nam, Biostatistics Branch, NCI, Executive Plaza South, Room 8028, 6120 Executive Boulevard, Rockville, MD 20892-7240, U.S.A.
[†] E-mail: namj@mail.nih.gov

Contract/grant sponsor: Intramural Research Program of NIH National Cancer Institute

Received 21 July 2005
Accepted 22 February 2006

[7] and Lantz and Nebenzahl [8] suggested the prevalence-adjusted, bias-adjusted κ, which is identical to the S coefficient [9, 10], and the G index of agreement [11]. It is also equivalent to the random error coefficient of agreement [12] and can be expressed as the difference between concordant and discordant rates. Another coefficient often used in the evaluation of agreement is the odds ratio. It is less dependent on marginals. Some authors, e.g. Spitznagel and Helzer [13] favoured Yule's index [14], i.e. the coefficient of colligation Y, which is based on the odds ratio.

We compare two correlated adjusted κ statistics and also two correlated Yule's Y's. For example, suppose that each subject is rated by three methods: two inexpensive and quick methods and the gold standard, which is very expensive and time consuming. We are interested in comparing the agreement between one inexpensive method and the gold standard *versus* that between the other inexpensive method and the gold standard to determine which of the inexpensive methods is preferable in terms of agreement with the gold standard. Statistical methods for the comparison using the adjusted agreement and Yule's Y have not been fully investigated.

In this paper, we review Cohen's κ, S coefficient of adjusted agreement and Yule's Y index. We examine their relations, and provide interval estimation of the adjusted agreement. In Section 3, we present the homogeneity test for correlated S statistics and interval estimation of the difference between two S coefficients. We also provide interval estimation of a ratio of odds ratios related to a comparison of Yule's Y where the odds ratios are correlated. We investigate relations of the coefficient of adjusted agreement, S, and Y index with sensitivity and specificity. Sections 4 and 5 contain numerical examples based on actual data and concluding remarks.

2. ADJUSTED AGREEMENT AND YULE'S Y INDEX

Consider the case where two raters classified each of N subjects according to whether an event is present (1) or absent (0). The N pairs of rating by raters A and G can be divided into four categories: $(1, 1)$, $(1, 0)$, $(0, 1)$, and $(0, 0)$. The outcome of N pairs of ratings and corresponding probabilities can be expressed in Table I.

2.1. *Adjusted agreement*

Let P_o and P_e denote the observed proportions of agreement and the proportion of agreement expected by chance. Note that P_o and P_e are estimates of the corresponding π values.

Table I. Observation and probabilities of raters A and G.

Observations					Probability model				
		Rater G					Rater G		
Rater	A	1	0	Sum	Rater	A	1	0	Sum
	1	a	b	m_1		1	π_{11}	π_{10}	$\pi_{1\bullet}$
	0	c	d	m_0		0	π_{01}	π_{00}	$\pi_{0\bullet}$
Sum		n_1	n_0	N	Sum		$\pi_{\bullet 1}$	$\pi_{\bullet 0}$	1

Cohen [1] introduced the κ coefficient as

$$\hat{\kappa} = (P_o - P_e)/(1 - P_e) = 1 - (1 - P_o)/(1 - P_e)$$

where $P_o = (a+d)/N$ and $P_e = (m_1 n_1 + m_0 n_0)/N^2$. Byrt *et al.* [7] defined the bias index (BI) and prevalence index (PI) as follows: BI is the difference in the proportion of positive for the two raters, i.e. $\mathrm{BI} = (b - c)/N$, and PI is the difference between the proportion of positive and the proportion of negative, i.e. $\mathrm{PI} = (a - d)/N$. To adjust κ for the difference in prevalence and for bias between two raters, b and c are replaced by their average, $(b+c)/2$, and also a and d by their average, $(a+d)/2$. The κ statistic based on the adjusted 2×2 table has been called the prevalence-adjusted, bias-adjusted κ, i.e.

$$\hat{S} = \left(\frac{a+d}{N} - \frac{1}{2}\right) \Big/ \left(1 - \frac{1}{2}\right) = 2P_o - 1$$

(e.g. Reference [7]) which is the estimated S coefficient [9] and random error coefficient [12]. It has been suggested as a reference standard for comparisons [8] and also called as Guilford's G index [11]. We refer to $S = 2(\pi_{11} + \pi_{00}) - 1$ as the adjusted κ or adjusted agreement throughout this paper. The adjusted κ, S, is a special case of $P_e = \frac{1}{2}$ in Cohen's κ. It is the difference between concordant and discordant proportions: $\hat{S} = \{a+d-(b+c)\}/N = 2(P_o - \frac{1}{2})$. We can show that $\hat{S} \geqslant \hat{\kappa}$ when $P_e \geqslant \frac{1}{2}$ and $\hat{S} < \hat{\kappa}$ when $P_e < \frac{1}{2}$. Suppose that matched samples are divided into either concordant or discordant pairs and NP_o is a binomial variable with parameters N and π where $\pi = \pi_{11} + \pi_{00}$. The variance of $\hat{S}$ is $\mathrm{var}(\hat{S}) = 1 - S^2$ and a simple interval estimate of S is

$$\hat{S} \pm z_{\alpha/2} \bullet (1 - \hat{S}^2)^{1/2} \tag{1}$$

where $z_{\alpha/2}$ is the $100(1 - \alpha/2)$ percentile point of the standard normal distribution. Using Wilson's score-like procedure [15], the improved confidence limits for a small sample size are

$$S = [\hat{S} \pm \{\hat{S}^2 - (1 + z_{\alpha/2}^2)(\hat{S}^2 - z_{\alpha/2}^2)\}^{1/2}]/\{(1 + z_{\alpha/2}^2)\} \tag{2}$$

The signs $-$ or $+$ are corresponding to the lower or upper limits. Consider that both marginals are fixed in Table I. The S statistic is rewritten as $\hat{S} = 2(2a + n_0 - m_1)/N - 1$. Since the variance of a can be expressed as $\mathrm{var}(a) = \sigma^2$ where $\sigma^2 = N/(\sum_{i=0}^{1} \sum_{j=0}^{1} \pi_{ij}^{-1})$ [16], we have $\mathrm{var}(\hat{S}) = (4/N)^2 \sigma^2$. A simple $100\%(1 - \alpha)$ confidence interval for S is

$$\hat{S} \pm 4 \bullet z_{\alpha/2} \bullet \hat{\sigma}/N \tag{3}$$

where $\hat{\sigma} = (1/a + 1/b + 1/c + 1/d)^{-1}$. For a small sample size, internal estimation of S using the score method may improve the accuracy for a confidence coefficient. The interval is obtained by solving the following equation with respect to π_{11}, i.e.:

$$(\hat{S} - S)^2/\mathrm{var}(\hat{S}) = z_{\alpha/2}^2 \tag{4}$$

where $S = 2(\pi_{11} + \pi_{00} - \frac{1}{2})$, $\pi_{10} = \pi_{1\bullet} - \pi_{11}$, $\pi_{01} = \pi_{\bullet 1} - \pi_{11}$, and $\pi_{00} = 1 - \pi_{11} - \pi_{10} - \pi_{01}$. Two permissible roots of equation (4) with respect to π_{11} are found numerically by an iterative procedure (Appendix A) and the limits of the interval for S follow. Since the variance of $\hat{S}$ under a multinomial model is smaller than that under a binomial model (Appendix B), we expect that the width of the interval estimation by (3) is shorter than that by (1).

2.2. Yule's index Y

The κ and S statistics range from -1 to 1, with their null values as 0, while the OR ranges from 0 to ∞, with its null value as 1. It is desirable that a measure based on OR has the same range as the κ or S so that the competing measure could easily be related back to these indices of agreement. Yule's coefficient of colligation is defined as $Y=(\psi^{1/2}-1)/(\psi^{1/2}+1)$ where $\psi=\pi_{11}\,\pi_{00}/(\pi_{10}\,\pi_{01})$. The estimator of Yule's coefficient is $\hat{Y}=\{(ad)^{1/2}-(bc)^{1/2}\}/\{(ad)^{1/2}+(bc)^{1/2}\}$ where $-1\leqslant\hat{Y}\leqslant 1$. Cohen's κ statistic is always smaller than $\hat{Y}$ (Appendix C). The estimated variance of $\hat{Y}$ is $\widehat{\text{var}}(\hat{Y})=(1-\hat{Y}^2)\hat{\sigma}^{-2}/16$ where $\hat{\sigma}^2=(1/a+1/b+1/c+1/d)^{-1}$, e.g. Reference [17]. The odds ratio based on the adjusted 2×2 table is $\text{OR}=\{(a+d)/2\}^2/\{(b+c)/2\}^2$. Thus, Yule's Y statistic for the adjusted 2×2 table is $\hat{Y}=\{a+d-(b+c)\}/N=2P_o-1$ which is $\hat{S}$. Cohen's κ, S and Y statistics are the same in the perfectly balanced 2×2 table. Since $\hat{Y}=1-2/(\text{OR}^{1/2}+1)$, Yule's Y statistic is monotonically increasing with respect to OR. Therefore, interval estimation of Y is also obtained by applying that of the odds ratio, e.g. using those of Woolf [18], Cornfield [19], Gart and Thomas [20].

The S statistic is based on an additive model involving four cell frequencies in the arithmetic scale, i.e. $\hat{S}=(a+d-b-c)/N$, while the logarithm of OR is based on an additive model for logarithmically transformed cell frequencies in the 2×2 tables, i.e. $\ln(\text{OR})=\ln(a)+\ln(d)-\ln(b)-\ln(c)$. The $\hat{S}$ and OR are conceptually similar but they are different in scales. The OR is sensitive to small frequencies of discordant pairs, b and c, and it is undefined when one of them is zero.

3. COMPARISON OF VALIDITY OF RATERS

In this paper, the term validity refers to agreement between the true state and the designating fallible rater. Suppose that each of N subjects is rated by methods A, B, and G where G is the gold standard. Results of ratings are summarized in Table II. The subscripts 1 and 0 could indicate exposed and not-exposed, or perhaps, positive and negative. Table I is a special case

Table II. Observation and probability of three rating methods (G: gold standard).

Rating method				
A	B	G	Observed number	Probabilities
1	1	1	x_{111}	π_{111}
1	0	1	x_{101}	π_{101}
0	1	1	x_{011}	π_{011}
0	0	1	x_{001}	π_{001}
1	1	0	x_{110}	π_{110}
1	0	0	x_{100}	π_{100}
0	1	0	x_{010}	π_{010}
0	0	0	x_{000}	π_{000}
Sum			N	1

of Table II. We are interested in comparing agreement between A and G with that between B and G, i.e. the validity of A and B.

3.1. Comparison of two adjusted κ's

Let a dot in subscripts denote the sum over 0 and 1, e.g. $\pi_{1\bullet 1} = \pi_{101} + \pi_{111}$. Also, S_1 and S_2 denote the coefficient of adjusted agreement for A and G and that for B and G, respectively, i.e. $S_1 = 2(\pi_{1\bullet 1} + \pi_{0\bullet 0}) - 1$ and $S_2 = 2(\pi_{\bullet 11} + \pi_{\bullet 00}) - 1$, and Δ will denote the difference between them, i.e. $\Delta = S_1 - S_2 = 2(\pi_{101} - \pi_{011} + \pi_{010} - \pi_{100})$ where $-1 \leqslant \Delta \leqslant 1$. The estimated value of Δ is $\hat{\Delta} = 2(x_{101} - x_{011} + x_{010} - x_{100})/N$, and its variance is $\text{var}(\hat{\Delta}) = 4\{(\pi_{101} + \pi_{011} + \pi_{010} + \pi_{100}) - (\pi_{101} - \pi_{011} + \pi_{010} - \pi_{100})^2\}/N$ (Appendix D). Thus, the standard error of $\hat{\Delta}$ may be written as $\text{SE}(\hat{\Delta}) = 2\{(x_{101} + x_{011} + x_{010} + x_{100}) - (x_{101} - x_{011} + x_{010} - x_{100})^2/N\}^{1/2}/N$, and an approximate $100\%(1 - \alpha)$ confidence interval for Δ is

$$\hat{\Delta} \pm z_{\alpha/2} \bullet \text{SE}(\hat{\Delta}) \tag{5}$$

When $\Delta = 0$, $\pi_{101} - \pi_{011} = \pi_{100} - \pi_{010}$ or $\pi_{101} + \pi_{010} = \pi_{100} + \pi_{011}$. The variance of $\hat{\Delta}$ under $\Delta = 0$ is $\{\text{var}(\hat{\Delta})\}_0 = 8T/N$ where $T = (\pi_{101} + \pi_{011} + \pi_{100} + \pi_{010})/2$. Since $\{\widehat{\text{var}}(\hat{\Delta})\}_0 = 8\hat{T}/N$ where $\hat{T} = (x_{101} + x_{011} + x_{010} + x_{100})/(2N)$, a statistic for testing $\Delta = 0$ against $\Delta \neq 0$ at level α is

$$X^2 = (x_{101} - x_{011} + x_{010} - x_{100})^2/(x_{101} + x_{011} + x_{010} + x_{100}) \tag{6}$$

We reject $\Delta = 0$ against $\Delta \neq 0$ at a when $X^2 \geqslant \chi^2_{1,\alpha}$ and do not reject otherwise. The $\chi^2_{1,\alpha}$ is the $100(1 - \alpha)$ percentile point of the χ^2 distribution with one degree of freedom. The test statistic (6) can be considered as an extension of McNemar test [21] to two 2×2 tables. It is a summary statistic for detecting a difference between A and B in both sensitivity and specificity.

3.2. Comparison of two Yule's Y statistics

Denote Yule's coefficient of colligation for A and G and that for B and G as Y_1 and Y_2, and denote a difference between them as $D = Y_1 - Y_2 = 2(\psi_1^{1/2} - \psi_2^{1/2})/\{(\psi_1^{1/2} + 1)(\psi_2^{1/2} + 1)\}$. Note that $Y_1 \geqslant Y_2$ as $\psi_1 \geqslant \psi_2$ and $Y_1 < Y_2$ as $\psi_1 < \psi_2$, or *vice versa*. Since $Y_1 = Y_2$ leads to $\psi_1 = \psi_2$ or $\ln \psi_1 = \ln \psi_2$, consider a difference between two log odds ratios, i.e. $\hat{R} = \ln(\text{OR}_1) - \ln(\text{OR}_2)$ where $\text{OR}_1 = x_{1\bullet 1}x_{0\bullet 0}/(x_{1\bullet 0}x_{0\bullet 1})$ and $\text{OR}_2 = x_{\bullet 11}x_{\bullet 00}/(x_{\bullet 10}x_{\bullet 01})$. Using the delta method, the asymptotic variance of $\hat{R}$ can be expressed as

$$\text{var}(\hat{R}) = \left[\sum_{k=0}^{1}\left\{\sum_{i=0}^{1}\pi_{i\bullet k}^{-1} + \sum_{j=0}^{1}\pi_{\bullet jk}^{-1} - \frac{2(\pi_{11k}\pi_{00k} - \pi_{10k}\pi_{01k})\pi_{\bullet\bullet k}}{\pi_{1\bullet k}\pi_{0\bullet k}\pi_{\bullet 1k}\pi_{\bullet 0k}}\right\}\right] \bigg/ N \tag{7}$$

(Appendix E). The third term on the inside of the second bracket is a measure of association for A and B, which is similar to the ϕ-coefficient. If there is no association for A and B, then this term is zero for $k = 0$ and 1. The Wald-type $100\%(1 - \alpha)$ confidence interval for R is

$$\hat{R} \pm z_{\alpha/2} \bullet \text{SE}(\hat{R}) \tag{8}$$

where $\{\text{SE}(\hat{R})\}^2$ is the asymptotic variance of $\hat{R}$, (7), evaluated at $\pi_{ijk} = x_{ijk}/N$, $\pi_{i\bullet k} = x_{i\bullet k}/N$, and $\pi_{\bullet jk} = x_{\bullet jk}/N$ for $i, j, k = 0, 1$. For extreme cases, $x_{i\bullet k} = 0$ and $x_{\bullet jk} = 0$ are replaced by

$x_{i\bullet k}=1/(2N)$ and $x_{\bullet jk}=1/(2N)$, respectively, so that the interval (8) is defined. Note that $R=\ln\psi_1-\ln\psi_2$ and $R=0$ leads to $\psi_1=\psi_2$ (or $Y_1=Y_2$) and *vice versa*. Grizzle *et al.* [22] presented a general approach to the analysis of categorical data by linear models. Under the framework of their approach with the proper constraint on parameters for a given problem, we can also derive the test statistic for comparing two adjusted κ's and Wald-type confidence interval for comparing Yule's Y indices.

3.3. Relations of S and Y indices with sensitivity and specificity

In assessment on validity of raters A and B, sensitivity and specificity are two useful measures. Since the rater G is the gold standard, the estimated sensitivity and specificity of rater A as $\hat{\mathrm{SEN}}_1=x_{1\bullet 1}/x_{\bullet\bullet 1}$ and $\hat{\mathrm{SPEC}}_1=x_{0\bullet 0}/x_{\bullet\bullet 0}$ and those for rater B are $\hat{\mathrm{SEN}}_2=x_{\bullet 11}/x_{\bullet\bullet 1}$, and $\hat{\mathrm{SPEC}}_2=x_{\bullet 00}/x_{\bullet\bullet 0}$. The S coefficient and Yule's Y are expressed as $S_i=2(p\bullet\mathrm{SEN}_i+q\bullet\mathrm{SPEC}_i)-1$ where $p=n_1/N$ and $q=n_0/N=1-p$ and $Y_i=(\mathrm{SEN}_i+\mathrm{SPEC}_i-1)/[(\mathrm{SEN}_i\bullet\mathrm{SPEC}_i)^{1/2}+\{(1-\mathrm{SEN}_i)(1-\mathrm{SPEC}_i)\}^{1/2}]^2$ for $i=1$ and 2. Both coefficients are monotone increasing functions of sensitivity and specificity. S is a sum of weighted sensitivity and specificity where weights are proportions of the positive and negative ratings by the gold standard. Y is some combination of sensitivity and specificity. The adjusted κ, S, and Yule's Y are global indices of agreement and useful in overall assessment on validity of A and B. The difference between the S coefficient for A and G and that for B and G is $S_1-S_2=2\{p\bullet(\mathrm{SEN}_1-\mathrm{SEN}_2)+q(\mathrm{SPEC}_1-\mathrm{SPEC}_2)\}$ where p and $q>0$, and that between Yule's Y's is proportional to $(\psi_1^{1/2}-\psi_2^{1/2})$. Therefore, we have

$$S_1\geqslant(<)S_2\quad\text{when}\ p\bullet(\mathrm{SEN}_1-\mathrm{SEN}_2)+q\bullet(\mathrm{SPEC}_1-\mathrm{SPEC}_2)\geqslant(<)0\qquad(9)$$

and

$$Y_1\geqslant(<)Y_2\quad\text{when}\ \left(\frac{\mathrm{SEN}_1}{1-\mathrm{SEN}_1}\right)\left(\frac{\mathrm{SPEC}_1}{1-\mathrm{SPEC}_1}\right)\geqslant(<)\left(\frac{\mathrm{SEN}_2}{1-\mathrm{SEN}_2}\right)\left(\frac{\mathrm{SPEC}_2}{1-\mathrm{SPEC}_2}\right)\qquad(10)$$

If both sensitivity and specificity of A are greater than those of B, then the adjusted agreement S and the Yule's Y for A and G are also greater than those for B and G from (9) and (10), respectively. Similarly, if both sensitivity and specificity of A are smaller than those of B, then the S and Y coefficients for A and G are also smaller than those for B and G. However, when rater A has a higher sensitivity but a lower specificity than rater B, or *vice versa*, results using S and Y indices may or may not be consistent. In this situation, it is difficult to determine preference between A and B unless the relative importance of sensitivity and specificity are given. Note that prevalence is involved in (9) but not directly in (10).

4. EXAMPLES

In a case–control study investigating the association between the asbestos-exposure and mesothelioma, the primary data come from interviews of the next-of-kin for patients with mesothelioma and for controls who died of other causes, excluding cancer, respiratory disease, suicide, and violence [23]. Three modes of exposure assessment were compared: the next-of-kin's response to a direct question regarding the asbestos-exposure of the subject, an assessment by a job-exposure matrix that relates specific jobs to exposure levels based on the National

Occupational Hazard Survey (NOHS) [24], and an exposure assessment based on review of the work histories by an occupational hygienist (expert). For these examples, we considered the assessment of the expert as the 'gold standard' which is expensive and time consuming. The observed asbestos-exposure ratings for 208 mesothelioma cases and 533 controls by the three exposure assessment methods are shown in Table III [25]. For each subject, the three asbestos-exposure assessment methods were applied. Statistical analyses are summarized in Tables IV and V. For cases, the κ agreement for next-of-kin respondent and expert exposure assessment

Table III. Observation of asbestos—exposure by three assessment methods.

Exposed (1) or unexposed (0)				
Respondent	NOHS	Expert	Number observed	
A	*B*	*G*	Cases	Controls
1	1	1	69	36
1	0	1	47	14
0	1	1	22	82
0	0	1	28	113
1	1	0	0	4
1	0	0	1	3
0	1	0	7	39
0	0	0	34	242
Total			208	533

Table IV. Agreement measured by κ, adjusted kappa S and Yule's Y.

		Cases		Controls	
Joint exposure categories		(A,G)	(B,G)	(A,G)	(B,G)
1	1	116	91	50	118
0	1	50	75	195	127
1	0	1	7	7	43
0	0	41	35	281	245
Total		208	208	533	533
Rate of agreement expected by chance		0.54	0.48	0.53	0.52
κ		0.47	0.24	0.19	0.34
(95% C.I.)[a]		(0.39, 0.58)	(0.14, 0.30)	(0.14, 0.22)	(0.27, 0.41)
Adjusted kappa S		0.51	0.21	0.24	0.36
(95% C.I.)[b]		(0.47, 0.55)	(0.13, 0.30)	(0.21, 0.28)	(0.29, 0.43)
Yule's Y		0.81	0.42	0.52	0.39
(95% C.I.)		(0.43, 0.89)	(0.22, 0.57)	(0.35, 0.64)	(0.30, 0.47)
Sensitivity		0.70	0.55	0.20	0.48
Specificity		0.98	0.83	0.98	0.85

[a] 95 per cent C.I. [4].

[b] 95 per cent C.I. by (3).

Table V. Agreement for respondent and expert *versus* agreement for NOHS and expert (comparison of respondent and NOHS in validity).

		Cases		Controls	
Respondent	NOHS	Expert = 1	Expert = 0	Expert = 1	Expert = 0
1	1	69	0	36	4
0	1	22	7	82	39
1	0	47	1	14	3
0	0	28	34	113	242
Total		166	42	245	288
$\Delta^{a} = 0$ *versus* $\Delta \neq 0$		$X^2 = 12.48$ $(p = 0.0004)$		$X^2 = 7.42$ $(p = 0.006)$	
$\hat{\Delta}^{b}$		0.298		-0.120	
(95% C.I.)		$(0.138, 0.458)$		$(-0.204, -0.036)$	
$\hat{R}^{c}$		2.752		4.999	
(95% C.I.)		$(0.518, 4.997)$		$(4.163, 5.835)$	

[a] $\Delta = S_1 - S_2$ where $-1 \leqslant \Delta \leqslant 1$.
[b] $\hat{\Delta} = \hat{S}_1 - \hat{S}_2$ where $-1 \leqslant \hat{\Delta} \leqslant 1$.
[c] $\hat{R} = \ln(\mathrm{OR}_1) - \ln(\mathrm{OR}_2)$ where $-\infty < \hat{R} < \infty$.

methods ($\hat{k}_1 = 0.47$) is two-fold stronger than that for the NOHS and expert ($\hat{k}_2 = 0.24$). When the κ statistics are adjusted, the former ($\hat{S}_1 = 0.51$) is even stronger than the latter ($\hat{S}_2 = 0.21$) and their difference is highly significant ($X_1^2 = 12.46$, $p < 0.001$) using (6). The 95 per cent confidence interval for the difference between the two adjusted κ's $(0.138, 0.458)$ by (5), does not contain zero. Similarly, Yule's index for respondent and expert ($\hat{Y} = 0.81$) is almost two times greater than that of NOHS and expert ($\hat{Y} = 0.42$), and the 95 per cent confidence interval for R, $(0.52, 4.99)$ by (8), does not cover zero. Both sensitivity and specificity of exposure assessment by respondent (0.70 and 0.98) are significantly higher than those by NOHS (0.55 and 0.83) using McNemar test. It is clearly demonstrated that the respondent method is better than NOHS in exposure assessment for cases.

For controls, the S statistic for respondent and expert ($\hat{S}_1 = 0.24$) is smaller than that for NOHS and expert ($\hat{S}_2 = 0.36$) and the difference between $\hat{S}_1$ and $\hat{S}_2$ is highly significant ($X_1^2 = 7.42$, $p = 0.006$) by (6), and the 95 per cent confidence interval for $\Delta = S_1 - S_2$ is $(-0.20, -0.04)$ by (5). The upper limit of the interval is smaller than zero. However, the odds ratio for respondent and expert is nearly twice greater than the odds ratio for NOHS and expert, and the corresponding Y values are $\hat{Y}_1 = 0.52$ and $\hat{Y}_2 = 0.39$, respectively, and the 95 per cent confidence interval for R is $(4.16, 5.84)$ by (8) and the lower limit is greater than zero. The preference of the NOHS method over respondent based on the adjusted agreement, S, is reversed when Yule's Y is applied. Results using S and Y coefficients are inconsistent for controls. In this example, the NOHS method provides a higher sensitivity than the respondent method (0.48 *versus* 0.20) but the former gives a lower specificity (0.85 *versus* 0.96). Both methods have very good specificity. However, the respondent method has very poor sensitivity. Overall, the NOHS method is better balanced and more sensible than the respondent for the controls. The sensitivity of the next-of-kin respondents for controls (20 per cent) is dramatically lower than that for cases (70 per cent). It is apparent that time dulled the memories of the respondents for controls far more than for cases.

The proportions of agreement expected by chance ranged from 0.48 to 0.54. Since they are close to $P_e = 0.5$, the values of κ are not greatly different from those of S coefficient. In this section, we illustrated two possible situations: a typical case where inference on comparing validity of assessment methods using S and that using Y are consistent, and an atypical situation where they are not consistent.

5. REMARKS

Cohen's κ is a popular index of agreement due to its simplicity and wide applicability. It is an overall measure of agreement and does not make a distinction among sources of disagreement. Since the value of Cohen's κ is influenced by both the prevalence and the presence of bias between raters, we may caution against its use for comparing raters in the strength of agreement when the marginal frequencies for κ statistics are quite different. It may be difficult to maintain consistent labels corresponding the values of κ statistics, e.g. labelling the ranges of κ statistics, <0, 0–0.20, 0.21–0.40, 0.41–0.60, 0.61–0.80, and 0.81–1.00 as poor, slight, fair, moderate, substantial, and almost perfect [26].

In the simultaneous comparison of both sensitivity and specificity of two binary diagnostic or screening tests, the equality of two test methods in sensitivity and specificity can be examined by the sum of two independent McNemar's statistics, e.g. Reference [27]. The hypothesis space is two-dimensional and the test statistic is a χ^2 with two degrees of freedom. However, rejection of the null hypothesis does not lead to a preference of one rating method over the other. It may be desirable to combine sensitivity and specificity so that it is possible us to make inference in the one-dimensional hypothesis. If the relative importance of sensitivity and specificity are given for a specific study, we can compare rating methods using a weighed mean of sensitivity and specificity where the weights are related to both the prevalence and the relative importance of sensitivity and specificity. When the sensitivity and specificity are equally important, the weighted mean is the same as the S coefficient. We recall that the statistic for testing the equality of two S statistics is a χ^2 with one degree of freedom (form (6), Section 3.1) while that by Lachenbruch and Lynch is a χ^2 with two degrees of freedom. The adjusted κ and Yule's Y index are conceptually similar and differ in scale. Both indices are reasonable measures, which combine sensitivity and specificity. It is difficult to make a judgment about the superiority of one index over the other. We suggest use of both indices in comparing assessment methods for validity. If the inference on comparing validity of two rating methods using adjusted κ and that using Yule's index are consistent, then it strengthens the conclusion. If they are not consistent, then it is also important to report the inconsistency and examine the characteristics of the rating methods.

In this paper, we considered two rating methods with the gold standard for each of N subjects and a comparison of validity of the rating methods. An alternative design might have N subjects rated by both rating method I and the gold standard as well as another N subjects rated by both the rating method II and the gold standard. However, such a design is inefficient and a wasteful use of the expensive gold standard. We suggest against this design in studies comparing the validity of rating methods. The design in Section 3 can be extended to the case of more than two rating methods with the gold standard and pairwise comparisons of validity of the rating methods.

In Section 3, we considered that one of three rating methods is the gold standard. However, in some agreement studies or diagnostic tests, it is impossible to know the true status or there is no the gold standard. For these cases, the statistical methods for comparing rating methods in Sections 3.1 and 3.2 can also be applied by replacing the gold standard with the reference standard, which is considered as more accurate than the other two methods. When there is no gold standard, it is inappropriate to use such terminologies as validity, sensitivity and specificity. We may evaluate statistical comparisons using the S coefficient of adjusted agreement and Yule's Y index together with the positive and negative predictive values.

APPENDIX A: INTERVAL ESTIMATION OF THE S COEFFICIENT BY THE SCORE METHOD

Rewrite (4) as

$$(\hat{S} - S)^2 - z_{\alpha/2}^2 \bullet \mathrm{var}(\hat{S}) = 0 \tag{A1}$$

where $\mathrm{var}(\hat{S}) = (4/N)^2\sigma^2$, $\sigma^2 = N/U$, $U = \sum_{i=0}^{1}\sum_{j=0}^{1} \pi_{ij}^{-1}$, $\pi_{10} = \pi_{1\bullet}$, $-\pi_{11}$, $\pi_{01} = \pi_{\bullet 1} - \pi_{11}$, $\pi_{00} = \pi_{11} + 1 - \pi_{1\bullet} - \pi_{\bullet 1}$, $S = 4\pi_{11} + 1 - 2(\pi_{1\bullet} + \pi_{\bullet 1})$ and $\hat{S} = 2P_{\mathrm{o}} - 1$. Using an approach similar to that of Cornfield [19] and Hale and Fleiss [4], we can obtain the limits of a $(1 - \alpha)$ confidence interval for S by finding π_{11} which satisfies the equation from (A1), i.e. $F \equiv (\hat{S} - S)^2 U - 4z_{\alpha/2}^2/N = 0$, using an iterative procedure, e.g. Newton–Ralphson's algorithm [28]. Denoting $V = (1/\pi_{11}^2 - 1/\pi_{10}^2 - 1/\pi_{01}^2 + 1/\pi_{00}^2)$ and $\pi_{11}^{(0)}$ as a trial value, the first iterated value of π_{11} is $\pi_{11}^{(1)} = \pi_{11}^{(0)} - F(\pi_{11}^{(0)})/F'(\pi_{11}^{(0)})$ where $F'(\pi_{11}^{(0)}) = -[(\hat{S} - S)\{8U + (\hat{S} - S)V\}]_{\pi_{11} = \pi_{11}^{(0)}}$. The procedure is repeated until the iterated values converge. Thus, the lower and upper limits of an interval estimation of S are found by the relation of $S = 4\pi_{11} + 1 - 2\ (\pi_{1\bullet} + \pi_{\bullet 1})$ corresponding to the lower and upper limits for π_{11}.

APPENDIX B: $\mathrm{var}(\hat{S}) < \mathrm{var}(\hat{S})_b$

From Section 2, the variances of $\hat{S}$ under multinomial and binomial models are expressed as

$$\mathrm{var}(\hat{S}) = \frac{4}{N}\left(\frac{4}{1/\pi_{11} + 1/\pi_{00} + 1/\pi_{10} + 1/\pi_{01}}\right)$$

and

$$\mathrm{var}(\hat{S})_b = 4\pi(1 - \pi)/N \quad \text{where } \pi = \pi_{11} + \pi_{00}$$

Since

$$\frac{1}{4}\left(\frac{1}{\pi_{11}} + \frac{1}{\pi_{00}}\right) - \frac{1}{\pi} = \frac{(\pi_{11} - \pi_{00})^2}{4\pi_{11}\pi_{00}\pi} > 0 \quad \text{and} \quad \frac{1}{4}\left(\frac{1}{\pi_{10}} + \frac{1}{\pi_{01}}\right) - \frac{1}{1 - \pi} = \frac{(\pi_{10} - \pi_{01})^2}{4\pi_{10}\pi_{01}(1 - \pi)} > 0$$

a summation of the two inequalities yields

$$\frac{1}{4}\left(\frac{1}{\pi_{11}}+\frac{1}{\pi_{00}}+\frac{1}{\pi_{10}}+\frac{1}{\pi_{01}}\right)>\frac{1}{\pi}+\frac{1}{1-\pi} \tag{B1}$$

The inverse of both sides of (B1) alters the direction of the inequality sign so $4/(\pi_{11}^{-1}+\pi_{00}^{-1}+\pi_{10}^{-1}+\pi_{00}^{-1})<\pi(1-\pi)$ which leads to $\mathrm{var}(\hat{S})<\mathrm{var}(\hat{S})_b$.

APPENDIX C: RELATIONSHIP BETWEEN ESTIMATED YULE'S Y AND κ STATISTIC

Since $\hat{Y}=\{(ad)^{1/2}-(bc)^{1/2}\}/\{(ad)^{1/2}+(bc)^{1/2}\}$ and $\hat{\kappa}=2(ad-bc)/(m_1n_0+m_0n_1)$ where $m_1=a+b$, $m_0=c+d$, $n_1=a+c$, and $n_0=b+d$, the κ statistic can be expressed as $\hat{\kappa}=2\{(ad)^{1/2}+(bc)^{1/2}\}^2\hat{Y}/(m_1n_0+m_0n_1)$ or $\hat{\kappa}/\hat{Y}=2\{(ad)^{1/2}+(bc)^{1/2}\}^2/(m_1n_0+m_0n_1)$. Since $m_1n_0+m_0n_1-2\{(ad)^{1/2}+(bc)^{1/2}\}^2=\{(ab)^{1/2}-(cd)^{1/2}\}^2+\{(ac)^{1/2}-(bd)^{1/2}\}^2+(b-c)^2>0$, we have $\hat{\kappa}/\hat{Y}<1$ or $\hat{\kappa}<\hat{Y}$.

APPENDIX D: VARIANCE OF $\hat{\Delta}$

The variance of $\hat{\Delta}=2(x_{101}-x_{011}-x_{100}+x_{010})/N$ is written as

$$\mathrm{var}(\hat{\Delta})=4\left[\sum_{k=0}^{1}\{\mathrm{var}(x_{10k})+\mathrm{var}(x_{01k})-2\,\mathrm{cov}(x_{10k},x_{01k})\}-2\{\mathrm{cov}(x_{101},x_{100})-\mathrm{cov}(x_{011},x_{100})-\mathrm{cov}(x_{101},x_{010})+\mathrm{cov}(x_{011},x_{010})\}\right]\Big/N^2$$

Since $\mathrm{var}(x_{ijk})=N\pi_{ijk}(1-\pi_{ijk})$ and $\mathrm{cov}(x_{ijk},x_{ijk'})=-N\pi_{ijk}\pi_{ijk'}$ for $k\neq k'$ and $i,j,k,k'=0,1$, a straightforward algebra yields

$$\mathrm{var}(\hat{\Delta})=4\{(\pi_{101}+\pi_{011}+\pi_{100}+\pi_{010})-(\pi_{101}-\pi_{011}-\pi_{100}+\pi_{010})^2\}/N$$

APPENDIX E: VARIANCE OF $\hat{R}$ WHERE $\hat{R}=\ln\hat{\psi}_1-\ln\hat{\psi}_2$

Define $R=\ln\psi_1-\ln\psi_2$ where ψ_1 and ψ_2 are the odds ratio for A and G and that for B and G, respectively. Consider the estimator of R as $\hat{R}=\ln\hat{\psi}_1-\ln\hat{\psi}_2$ where $\hat{\psi}_1$ and $\hat{\psi}_2$ are correlated. Applying a series expansion of $\hat{R}$ with respect to x_{ijk} for $i,j,k=0.1$, the asymptotic variance of $\hat{R}$ is expressed as

$$\mathrm{var}(\hat{R})=\left[\sum_{i=0}^{1}\sum_{j=0}^{1}\sum_{k=0}^{1}\pi_{ijk}\left(\frac{\partial\hat{R}}{\partial x_{ijk}}\right)^2_{x_{ijk}=N\pi_{ijk}}-\left\{\sum_{i=0}^{1}\sum_{j=0}^{1}\sum_{k=0}^{1}\pi_{ijk}\left(\frac{\partial\hat{R}}{\partial x_{ijk}}\right)_{x_{ijk}=N\pi_{ijk}}\right\}^2\right]$$

We can show that the sum of terms inside of the second bracket of the above equation is zero. After straightforward algebraic operations, the variance of $\hat{R}$ can be written as

$$\operatorname{var}(\hat{R}) = \left[\sum_{k=0}^{1}\left\{\sum_{i=0}^{1}\pi_{i\bullet k}^{-1}+\sum_{j=0}^{1}\pi_{\bullet jk}^{-1}-\frac{2(\pi_{11k}\pi_{00k}-\pi_{10k}\pi_{01k})\pi_{\bullet\bullet k}}{\pi_{1\bullet k}\pi_{0\bullet k}\pi_{\bullet 1k}\pi_{\bullet 0k}}\right\}\right]\Bigg/ N$$

ACKNOWLEDGEMENTS

The authors thank the reviewers for their comments to improve the presentation of this paper. This research was supported by the Intramural Research Program of the NIH National Cancer Institute.

REFERENCES

1. Cohen J. A coefficient of agreement for nominal scales. *Education and Psychological Measurement* 1960; **20**:37–46.
2. Fleiss JL, Cohen J, Everitt BS. Large-sample standard errors of kappa and weighted kappa. *Psychological Bulletin* 1969; **72**:323–327.
3. Fleiss JL, Cicchetti DV. Inference about weighted kappa in the non-null case. *Applied Psychological Measurement* 1978; **2**:113–117.
4. Hale C, Fleiss J. Interval estimation under two study designs for kappa with binary classification. *Biometrics* 1993; **49**:523–534.
5. Grove VM, Adreason NC, McDonald-Scott P, Keller MB, Shapiro RW. Reliability studies of psychiatric diagnosis. *Archives of General Psychiatry* 1981; **38**:408–413.
6. Feinstein AR, Cicchetti DV. High agreement but low kappa. I. The problems of two paradoxes. *Journal of Clinical Epidemiology* 1990; **43**:543–549.
7. Byrt T, Bishop J, Carlin JB. Bias prevalence and kappa. *Journal of Clinical Epidemiology* 1993; **46**:423–429.
8. Lantz CA, Nebenzahl E. Behavior and interpretation of the κ statistic: resolution of the two paradoxes. *Journal of Clinical Epidemiology* 1996; **49**:431–434.
9. Bennett EM, Albert R, Goldstein AC. Communications through limited response questioning. *Public Opinion Quarterly* 1954; **18**:303–308.
10. Zwick R. Another look at interrater agreement. *Psychological Bulletin* 1988; **103**:374–378.
11. Holly JW, Guilford JP. A note on the G-index of agreement. *Educational and Psychological Measurement* 1964; **XXIV**(4):749–753.
12. Maxwell AE. Coefficients of agreement between observers and their interpretation. *British Journal of Psychiatry* 1977; **130**:79–83.
13. Spitznagel EL, Helzer JE. A proposed solution to the base rate problem in the kappa statistic. *Archives of General Psychiatry* 1985; **42**:725–728.
14. Yule GU. On the methods of measuring association between two attributes. *Journal of the Royal Statistical Society* 1912; **75**:581–642.
15. Wilson EB. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association* 1927; **22**:209–212.
16. Stevens WL. Mean and variance of an entry in a contingency table. *Biometrika* 1951; **38**:468–470.
17. Bishop YMM, Fienberg SE, Holland PW. *Discrete Multivariate Analysis*: *Theory and Practice*. The MIT Press: Cambridge, MA, 1975.
18. Woolf B. On estimating the relationship between blood group and disease. *Annals of Human Genetics* 1955; **19**:251–253.
19. Cornfield J. A statistical problem arising from retrospective studies. In *Proceedings of the Third Berkeley Symposium IV*, Neyman J (ed.). University of California Press: Berkeley, CA, 1956; 133–148.
20. Gart JJ, Thomas DG. Numerical results on approximate confidence limits for the odds ratio. *Journal of the Royal Statistical Society*, *Series B* 1972; **34**:441–447.
21. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 1947; **12**:153–157.
22. Grizzle JE, Starmer CF, Koch GG. Analysis of categorical data by linear models. *Biometrics* 1969; **25**:489–504.
23. Spirtas R, Heineman EF, Berstein L, Beebe GW, Keehn RJ, Stark A, Harlow BL, Benichou J. Malignant mesothelioma: attributable risk of asbestos exposure. *Occupational and Environmental Medicine* 1994; **51**:804–811.

24. Seiber KW, Sundin DS, Frazier TM, Robinson CF. Development, use and availability of a job exposure matrix based on National Occupational Hazard Survey Data. *American Journal of Industrial Medicine* 1991; **20**: 163–174.
25. Nam J, Rice C, Gail MH. Comparison of asbestos exposure assessments by next-of-kin respondents, by an occupational hygienist, and by a job-exposure matrix from the National Occupational Hazard Survey. *American Journal of Industrial Medicine* 2005; **47**:443–450.
26. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; **33**:159–174.
27. Lachenbruch PA, Lynch CJ. Assessing screening tests: extension of McNemar's test. *Statistics in Medicine* 1998; **17**:2207–2217.
28. Cheney W, Kincaid D. *Numerical Mathematics and Computing*. Brooks/Cole Publishing Company: Belmont, CA, 1985.